华为“韬定律”到底是什么？

本报记者 姬晓婷

近日，国际电路与系统研讨会在上海举行。华为公司董事、半导体业务部副总裁何庭波发表演讲，正式发表“韬(τ)定律”。随后，何庭波关于“韬(τ)定律”的系统阐释文章《A Time Scaling Theory for Multi-Layer Electronic Systems（多层电子系统的时间缩放理论）》发表在中国科学院科技论文预发布平台。研读完这篇文章，记者发现：“韬(τ)定律”的提出将给中国半导体产业格局带来颠覆性的影响。

图为华为公司董事、半导体业务部副总裁何庭波在国际电路与系统研讨会上演讲

“韬(τ)定律”提出源自摩尔定律放缓

在过去大半个世纪里，全球半导体产业就像坐在一辆由“摩尔定律”驾驶的高速列车上。列车的换挡油门很简单：把晶体管做得更小、更密。摩尔定律的这句话业内人士估计都听到耳朵起茧子了——当价格不变时，集成电路上可容纳的晶体管数目，约每隔18~24个月便会增加一倍，性能也将提升一倍。

而这套指导了半导体产业半个多世纪的“旧时代”规则正在放缓。何庭波在论文的开篇便一针见血地指出了当前全球半导体行业面临的三个残酷现实：一是物理收益趋平。晶体管的尺寸已经快缩小到原子的物理极限了，现在费尽九牛二虎之力把它做小，性能的提升却微乎其微；二是“钞能力”开始失效。以前研发新工艺省钱，现在设计一颗最先进的顶尖芯片，研发预算居然要超过10亿美元，研发、采用新工艺的性价比正在降低；三是越小反而越贵。过去是工艺越先进，单个晶体管越便宜，现在倒过来了，工艺越先进，单个晶体管的造价不降反升。

对于中国半导体产业而言，在无法轻易拿到最先进光刻机的现实下，这个几何尺寸的“物理墙”来得更早、更猛烈。

“韬（T）定律”的提出，从根本上是为了解决一个核心问题：摩尔定律面临失效，整个产业该以什么为指导性能提升的核心指标？

为回答这个问题，何庭波选择了一条以时间为指标的优化道路。她提出，未来衡量芯片和系统进步的首要指标，不再是“它是几纳米”，而是“它完成一项任务需要多少时间。这就是τ缩放定律（简称“韬定律”），其中τ是物理学中代表“时间常数”的希腊字母。

“韬定律”优越性何在？

要成为指导全行业发展的一条定律，首先要实现的，便是定律的普适性。而“韬定律”，是一个横跨了“十二个数量级”的全能总指挥棒。

如果把计算基础设施从微观到宏观剖解开，那么一套计算基础设施大致可分成晶体管级、电路层级、芯片级、系统/数据中心级几个层次。而一个任务的完成时间与每个层级的性能和表现都息息相关。

这就是“韬定律”的魅力所在，原本芯片性能代际提升只是晶体管层的事情，整个半导体产业的发展决定权在于台积电这样的制造企业。而现在，在这套追求更低任务完成时间的评价体系中，每一个层级的从业者/部门，拥有了同样的任务，那就是更快。

而这个时间常数“τ”，在不同层级所表示的数字不一样。最微观的底层，它可以是单个晶体管切换一次电信号所花的几皮秒；在最中间的芯片层，它可以是处理器去缓存拿数据所花的几纳秒或微秒；在最宏观的顶层，它可以是整个AI数据中心协同处理完一个超大模型计算任务并做出响应所花的几秒钟。

以前，芯片产业链是个“冷漠社区”。做代工的只管把晶体管做小，画电路图的只管布线，做软件系统的只管写代码，大家语言不通。现在，τ定律强行把所有人拉到同一个账本前：全部用时间单位来算账。工艺专家省下的5皮秒，和架构师、软件专家省下的5皮秒，在总账本里的权重一模一样。

在这样的规则牵引下，产业链的运作方式将转变为哪一层堵，我们就在哪一层给系统“挤时间”。这种用“时间”定义“性能”的全新全局思维，就是τ定律的核心灵魂。

“韬定律”如何帮助芯片变强？

盯着“时间变短”，怎么就能让芯片性能变强呢？

摩尔定律与芯片性能之间的关系很容易理解。晶体管更小了，芯片单位面积内晶体管更多了，自然性能更强了。而现在，在“韬定律”的指导下，为什么更少的任务运行时间，会对应着产品性能更强的结果呢？

这是个数学问题。举个例子，我们平时买手机、买服务器，CPU核心频率是一个非常重要的判断芯片性能的指标。

而刚刚我们讨论的特征时间常数“τ”，拆解到了芯片系统的四大核心层级中，便对应着很多具体的指标。在晶体管层（τ_transistor-皮秒级）：它代表单个晶体管由“开”到“关”切换一次信号的时间；在电路层（τ_circuit-纳秒级）：它代表信号在微小的金属导线和逻辑门之间传递与充电消耗的时间，即所谓的RC传播延迟；在芯片层（τ_chip-微秒级）：它代表计算核心去旁边的缓存（SRAM）或大内存里搬运数据所花的时间；在系统/数据中心级（τ_system-秒级）：它成千上万颗芯片通过光纤、网络互相通信，协同处理完一个庞大AI任务的响应时间。

芯片的时钟频率（比如3.1GHz），代表芯片的“心脏”一秒钟能跳动31亿次。心脏跳一次的周期（处理一步计算的时间），就是频率的倒数。

在设计芯片时，频率往往被一条最慢、最长的导线卡死，这条路在工程上叫“关键路径（Critical Path）”。不管别的路线跑得多快，时钟必须等这条最慢的路走完，才能进行下一次跳动。

如果我们通过架构调整，把这条最慢路径上的信号通过时间，也就是把电路层的特征时间常数τ_circuit狠狠压缩，那芯片的“心脏”就不需要等那么久了，它就可以跳得更快。

如何降低时间常数τ？

在讲清楚基本原理之后，何庭波花了很大的篇幅来讲述如何降低时间常数τ。其中包括：LogicFolding（逻辑折叠技术）、统一总线（Unified Bus）、Hi-ONE近封装光学I/O技术等。这些技术其实是近几年业界讨论和实践最多的，简单解释一下：

首先是LogicFolding，以前的传统芯片（平房），所有的晶体管和逻辑门都平铺在一个二维平面上，只有最底部一层是具有计算功能的“激活层Active Tier”。如果两个逻辑门在平面上隔得远，中间就得连一根很长的金属导线。导线一长，寄生电阻和电容（RC延迟）就大，电信号走得慢，还特别耗电。

因此，华为放弃了平面假设。他们把芯片“折叠”起来，盖成了多层的三维楼房。原本在平面上相隔很远的两个逻辑门，被重新安排，一个放“一楼”，另一个直接放它头顶的“二楼”。

“一楼”和“二楼”之间，通过混合键合（Hybrid Bonding）技术实现“通信”。这种技术要求把两片晶圆的表面磨得像镜子一样平，达到原子级平整度，然后让上下的铜接点实现分子级的融合。通过超细微间距的混合键合，在上下两层芯片之间打通无数个垂直的“电梯通道”。因为信号从一楼到二楼走的是“垂直电梯”，物理距离缩短了30%以上。导线变短，电阻和电容暴跌，电路层的时间常数τ_circuit被强行压缩。

论文中以Kirin 2026（麒麟2026芯片）为例，在工艺节点完全没变的情况下，单位平面面积内的晶体管数量从1.55亿直接拉升到了2.38亿每平方毫米，大幅提升55%。

其次是统一总线。在传统体系下，服务器A的芯片要跟服务器B的芯片聊天，数据要跨越PCIe总线、打包成网络协议、走光纤、再解包，好比跨国运货还要办签证，效率极低。

华为的Unified Bus引入了“内存语义（Memory-semantic）”。简单来说，就是打破服务器之间的“行政主权边界”。在系统眼里，整个集群几万颗芯片的内存被拍平了，共享同一个物理地址空间。隔壁服务器的内存，芯片要拿数据直接去读写物理地址即可，连协议包装都省了。

这一改，跨节点获取数据的时间从过去的几十微秒，直接暴跌到了150纳秒以下。多台分离的服务器在逻辑上被合并成了一颗“巨大的虚拟单体芯片”。

然后是Hi-ONE近封装光学I/O。数据量太大时，传统的铜导线传电信号不仅发热恐怖，而且衰减严重。华为的解法是“电退光进”。他们研发了Hi-ONE技术，把微型的硅光子收发器（把电变成光、光变成电的零件）直接贴在AI核心芯片的家门口。数据一出计算核心，立刻变成一束激光通过光纤射出去。

最后是边缘至表面3D折叠（Edge-to-Surface 3D Folding）。主板上不再是平铺芯片，而是像玩“俄罗斯方块”一样，在三维立体空间里将加速芯片、存储芯片、光通信模块进行疯狂的纵向堆叠与嵌套，让彼此靠得更近，将空间距离压榨到极限。

通过这些技术的联合轰炸，何庭波在论文中预测，到2035年，AI硬件系统的集成度（在特定体积内发挥出的算力和存储密度）将实现100倍以上的增长。

τ原生EDA工具链成为关键

看到这里，很多人可能会热血沸腾，觉得我们马上就能实现反超了。但我们必须看到硬币的另一面：要让τ定律方案真正转起来，产业链需要经历一场痛苦的重构。

其中最难啃的骨头，就是论文中提到的EDA（电子设计自动化）工具链。以往设计芯片的软件工具（EDA）都是在二维孤岛下运行的。团队A负责平面布线，画完交给团队B，最后交给团队C去算散热。如果团队C发现几层晶体管叠在一起发热太厉害，会把芯片烧煳，那么整个项目将可能面临推倒重来的风险。

而在τ定律时代，这种生产方式将从原生设计自动化环境开始重构。这个新型工具链最大的特色是“跨层三维空间协同优化”。也就是说，工程师在软件里画下第一笔电路时，软件就会在三维空间里同时计算三件事：电路怎么走信号最快（电学约束）、怎么叠最不容易烧坏（热学物理场）、这样的硬件层配什么样的大模型算法最省时间（算法约束）。

不仅如此，产业链的合作方式也将迎来重构，芯片设计企业、代工企业和封装企业将走向“全栈一体化融合”。系统厂商在刚提出大模型需求时，就得把芯片设计商、封装厂、设备商叫到一张桌子上，共享底层的物理和热学参数，联合设计。

中国芯片产业的战略突围宣言

西方在几何缩放（光刻机）路线上跑了60年，构筑了极其坚固的专利和设备壁垒。如果我们一味地在别人的赛道上死磕“1纳米、2纳米”，不仅面临难以逾越的设备大山，更是在用自己的短板去硬碰别人的长板。

何庭波在2026年发表的这篇τ缩放理论，实际上是中国半导体产业的一份“自立新路线的独立宣言”。它明确告诉全行业：当几何尺寸的红利到头，或者路被堵死的时候，我们完全可以用“系统工程的整合能力”去对冲、战胜“单一单体芯片的工艺短板”。以时空换几何，以系统赢单点。

对于习惯了传统2D芯片设计的工程师和资本而言，这场向3D、向跨层协同的转型充满了痛苦和未知。这些愿景的实现也需要产业界上下游的共同参与。就像何庭波在文章最后所提到的那样：“大量开放问题，无单一组织可独立解决——工具链、标准、基准、器件物理、经济模型均需跨界协作。”

“本文既是一线实践报告，也是产业邀请。前路充满挑战，但方向明确无误。”何庭波在论文中指出。

券商评“韬定律”：EDA、晶圆代工、制造设备受影响最大

华为提出的“韬定律”引发产业界热议。5月26日，华创证券研究所电子组研究员张文瑶在接受《中国电子报》记者采访时指出：EDA、晶圆代工、制造设备是在“韬定律”路径指引下受影响最大的三个环节。

5月25日，华为公司董事、半导体业务部总裁何庭波在2026国际电路与系统研讨会上正式发表“韬（τ）定律”。“韬定律”提出的背景是摩尔定律即将走到极限，在物理层面继续把芯片做小、做精密变得越来越困难。在这种情况下，华为提出的“韬定律”实际上是一种“时间缩微”的概念，即把尽可能压缩信号传播时延作为芯片优化的核心指标。

在张文瑶看来，“韬定律”的提出对国内半导体产业链的影响可以从设计、制造、设备三个层面来看：

一是芯片设计环节。“韬定律”的指引下，芯片设计的方法需要从2D全面转成3D，EDA工具链将因此被重构。

二是晶圆代工环节。何庭波描述的“韬定律”实现路径中的“逻辑折叠”，需要在晶圆代工环节完成。国内现有晶圆代工企业工艺发展至14纳米、7纳米水平，若通过逻辑折叠提升效能，单颗芯片便能满足更多AI、手机等高端芯片需求，这将使国内代工厂产能更饱和、需求更旺盛，面向的市场空间也会更大。

三是制造设备。“韬定律”的实现需要电路板精度更高的先进封装堆叠技术，包括混合键合、TSV打孔、电镀设备、CMP等在内的制造设备都需要全面升级，精度要求远超以往。

对于“韬定律”的态度，境内外厂商存在明显差异。张文瑶表示，中国大陆境内的企业对于“韬定律”的态度大多比较积极，而包括中国台湾在内的境外企业，对“韬定律”的态度相对谨慎。这些境外公司认为，“韬定律”工程化落地的难度太大，工程实现过程中对精度、混合键合、平整度、散热的要求太高；实现“韬定律”对整个产业链的配合程度的要求也非常高，产业链中只要有一个环节“掉链子”，整个项目就无法落地。这些都是境外企业对“韬定律”持观望态度的重要原因。而且，由于海外企业能够拿到EUV光刻机或先进制程产能，可以持续走晶体管微缩路线，因此也就没有探索“韬定律”这一新路径的动力。

“韬定律”是一套工程性升级的方案，它的落实需要产业链诸多环节的配合，当前，产业链上的企业大多处于学习阶段。“‘韬定律’能否真正落地、能否重新定义行业，仍需要一段时间来观察。”张文瑶说道。

（姬晓婷）

第一季度全球DRAM/NAND Flash市场规模同比增长245%

本报讯 近日，CFM闪存市场发布的《2026年第一季度DRAM/NAND Flash营收市占排名》中显示，2026年第一季度，全球存储行业进入AI驱动的结构性景气上行周期，市场依然由卖方主导。需求端，AI算力建设及Agentic AI应用拉动数据中心长期供货协议（LTA）陆续落地，HBM、服务器DRAM与企业级SSD优先抢占产能，消费及移动端需求被动受压。供给端，原厂持续优化产品结构，向高附加值AI存储倾斜，而新增有效产能释放周期较长，短期供需缺口持续存在。

当前，DRAM与NAND合约价格环比大幅上行，原厂议价能力显著提升，紧缺涨价格局仍具备较强延续性。据CFM闪存市场分析，2026年第一季度全球DRAM/NAND Flash市场规模达1371.4亿美元，环比增长81.6%，同比增长245%，再创历史季度新高。其中，全球DRAM市场规模943.25亿美元，环比增长81.5%。数据中心需求激增已导致DRAM供应紧张，HBM占据大部分产能进一步挤占其他DRAM供给，推动第一季度DDR/LPDDR产品ASP大幅上涨，目前通用型DDR利润率已超过HBM。

排名方面，三星、SK海力士、美光仍占据前三甲，第一季度市场份额分别为40.5%、29.6%、19.9%。长鑫存储市场份额提升至7.7%，排名第四。

随着企业级SSD需求倍增并仍在加速，尽管消费类需求现疲软迹象，但不改整体供不应求的局面，2026年第一季度全球NAND Flash市场规模为428.15亿美元，环比增长81.8%。

排名方面，三星、SK海力士、铠侠、闪迪、美光位列前五，第一季度市场份额分别为29.7%、17.6%、14.9%、13.9%、11.7%。

（幸 稳）

Rambus推出DDR5 9600客户端内存模块芯片组

本报讯 记者**许子皓**报道：近日，Rambus宣布推出完整的DDR5 9600客户端内存模块芯片组，专为新一代AI PC中的高性能CUDIMM、CQDIMM和CSODIMM模块设计。该芯片组包含全新的第二代客户端时钟驱动器，支持9600MT/s的PC内存模块运行速度，并配备电源管理IC和串行存在检测集线器，从而带来更好的性能表现。

随着代理式人工智能的兴起，PC现已能够实时规划、执行和调整工作流程。这些工作负载需要持久的上下文、并行处理以及处理器与系统内存之间的持续数据传输，这要求带宽和容量均需大幅提升。与此同时，将DDR5内存频率提升至6400MT/s以上带来了新的技术挑战，包括信号衰减、时钟抖动和时序不稳定。

IDC研究副总裁Jeff Janukowicz表示：“为满足日益增长的性能需求，业界正向CUDIMM和CSODIMM等带时钟的内存架构转型，这些架构旨在解决更高数据速率下的信号完整性和时序挑战。”

全新的Rambus DDR5 9600客户端芯片组为工作频率在8000至9600MT/s之间的带时钟DDR5模块提供了完整的解决方案。该芯片组专为高性能和可扩展性而设计，支持新一代AI个人电脑、笔记本电脑和工作站。通过在模块层面解决信号完整性、供电和系统协调问题，Rambus简化了高性能内存模块解决方案的设计和部署。

Rambus内存接口芯片业务高级副总裁兼总经理Rami Sethi表示：“DDR5 9600客户端芯片组搭载了第二代客户端时钟驱动器，为新时代的智能高性能客户端系统奠定性能基础，助力AI驱动的生产力、新一代游戏以及专业内容创作。”

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号